Vision Research 49 (2009) 1295–1306

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Bayesian surprise attracts human attention

Laurent Itti [a,*], Pierre Baldi [b,1]

[a] Computer Science Department and Neuroscience Graduate Program, University of Southern California, Hedco Neuroscience Building, 3641 Watt Way, HNB-30A, Los Angeles, CA 90089, USA
[b] Computer Science Department and Institute for Genomics and Bioinformatics, University of California, Irvine, Irvine, CA 92697-3425, USA

ARTICLE INFO

Article history:
Received 3 October 2007
Received in revised form 2 September 2008

Keywords:
Attention
Surprise
Bayes theorem
Information theory
Eye movements
Natural vision
Free viewing
Saliency
Novelty

ABSTRACT

We propose a formal Bayesian definition of surprise to capture subjective aspects of sensory information. Surprise measures how data affects an observer, in terms of differences between posterior and prior beliefs about the world. Only data observations which substantially affect the observer's beliefs yield surprise, irrespectively of how rare or informative in Shannon's sense these observations are. We test the framework by quantifying the extent to which humans may orient attention and gaze towards surprising events or items while watching television. To this end, we implement a simple computational model where a low-level, sensory form of surprise is computed by simple simulated early visual neurons. Bayesian surprise is a strong attractor of human attention, with 72% of all gaze shifts directed towards locations more surprising than the average, a figure rising to 84% when focusing the analysis onto regions simultaneously selected by all observers. The proposed theory of surprise is applicable across different spatio-temporal scales, modalities, and levels of abstraction.

1. Introduction and background

In a world full of surprises, animals have developed an exquisite ability to quickly detect and orient towards unexpected events (Ranganath & Rainer, 2003). Yet, at present, our formal understanding of what makes an observation surprising is limited: Indeed, our everyday vocabulary lacks a quantitative notion of surprise, with qualities such as "wow factors" still ill-defined and thus far intractable to quantitative analysis. Here, within the Bayesian probabilistic framework, we develop a simple quantitative theory of surprise. Armed with this theory, we provide direct experimental evidence that Bayesian surprise best characterizes what attracts human gaze in large amounts of natural video stimuli.

Our effort to formally and mathematically define surprise is motivated by the fact that informal correlates of surprise have been described at nearly all stages of neural processing. Thus, surprise is an essential concept for the study of the neural basis of behavior. In sensory neuroscience, for example, it has been suggested that only the unexpected at one stage of processing is transmitted to the next stage (Rao & Ballard, 1999). Hence, sensory cortex may have evolved to adapt to, to predict, and to quiet down the expected statistical regularities of the world (Olshausen & Field, 1996; Müller, Metha, Krauskopf, & Lennie, 1999; Dragoi, Sharma, Miller, & Sur, 2002; David, Vinje, & Gallant, 2004), focusing instead on events that are unpredictable or surprising (Fairhall, Lewen, Bialek, & de Ruyter Van Steveninck, 2001). Electrophysiological evidence for this early sensory emphasis onto surprising stimuli exists from studies of adaptation in visual (Maffei, Fiorentini, & Bisti, 1973; Movshon & Lennie, 1979; Müller et al., 1999; Fecteau & Munoz, 2003), olfactory (Kurahashi & Menini, 1997; Bradley, Bonigk, Yau, & Frings, 2004), and auditory cortices (Ulanovsky, Las, & Nelken, 2003), subcortical structures like the LGN (Solomon, Peirce, Dhruv, & Lennie, 2004), and even retinal ganglion cells (Smirnakis, Berry, Warland, Bialek, & Meister, 1997; Brown & Masland, 2001) and cochlear hair cells (Kennedy, Evans, Crawford, & Fettiplace, 2003): neural responses greatly attenuate with repeated or prolonged exposure to an initially novel stimulus. At higher levels of abstraction, surprise and novelty are also central to learning and memory formation (Ranganath & Rainer, 2003), to the point that surprise is believed to be a necessary trigger for associative learning (Schultz & Dickinson, 2000; Fletcher et al., 2001), as supported by mounting evidence for a role of the hippocampus as a novelty detector (Knight, 1996; Stern et al., 1996; Li, Cullen, Anwyl, & Rowan, 2003). Finally, seeking novelty is a well-identified human character trait, possibly associated with the dopamine D4 receptor gene (Ebstein et al., 1996; Benjamin et al., 1996; Lusher, Chandler, & Ball, 2001).

Empirical and often *ad-hoc* formalizations of surprise, usually referred to as spatial "saliency" or temporal "novelty," are at the core of many laboratory studies of attention and visual search: The strongest attractors of attention are stimuli that pop-out from

* Corresponding author. Fax: +1 213 740 5687.
E-mail addresses: itti@usc.edu (L. Itti), pfbaldi@ics.uci.edu (P. Baldi).
[1] Tel.: +1 949 824 5809; fax: +1 949 824 4056.

0042-6989/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2008.09.007

their neighbors in space or time, like a salient vertical bar embedded within an array of horizontal bars (Treisman & Gelade, 1980; Wolfe & Horowitz, 2004), or the abrupt onset of a novel bright dot in an otherwise empty display (Theeuwes, 1995). Computationally, these notions may be summarized in terms of outliers (Markou & Singh, 2003) and Shannon information: stimuli which have low likelihood given a distribution of expected or learned stimuli, over space or over time, are outliers, are more informative in Shannon's sense, and capture attention (Duncan & Humphreys, 1989). We show that this line of thinking at best captures an approximation to surprise, but can be flawed in some extreme cases. To exacerbate the differences and to gauge their practical impact in ecologically relevant situations, we quantitatively compare Bayesian surprise to 10 existing measures of saliency and novelty, in their ability to predict human gaze recordings on large amounts of natural video data. We find that Bayesian surprise best characterizes where people look, even more so for stimuli that are consistently fixated by multiple observers. Our results suggest that surprise is an important formalization for understanding neural processing and behavior, and is the best known attractor of human attention.

This work extends Itti and Baldi (2006), through a more complete exposition of the theory and of the new proposed unit of surprise (the "wow"), simple examples of how surprise may be computed, and a broader set of experiments and comparisons with competing theories and models.

2. Theory

In this paper, we elaborate a definition of surprise that is general, information-theoretic, derived from first principles, and formalized analytically across spatio-temporal scales, sensory modalities, and, more generally, data types and data sources. Two elements are essential for a principled definition of surprise. First, surprise can exist only in the presence of uncertainty. Uncertainty can arise from intrinsic stochasticity, missing information, or limited computing resources. A world that is purely deterministic and predictable in real-time for a given observer contains no surprises. Second, surprise can only be defined in a relative, subjective, manner and is related to the expectations of the observer, be it a single synapse, neuronal circuit, organism, or computer device. The same data may carry different amounts of surprise for different observers, or even for the same observer taken at different times.

2.1. Defining surprise

In probability and decision theory it can be shown that, under a small set of axioms, the only consistent way for modeling and reasoning about uncertainty is provided by the Bayesian theory of probability (Cox, 1964; Savage, 1972; Jaynes, 2003). Furthermore, in the Bayesian framework, probabilities correspond to subjective degrees of beliefs in hypotheses (or so-called models). These beliefs are updated, as data is acquired, using Bayes' theorem as the fundamental tool for transforming prior belief distributions into posterior belief distributions. Therefore, within the same optimal framework, a consistent definition of surprise must involve: (1) probabilistic concepts to cope with uncertainty and (2) prior and posterior distributions to capture subjective expectations. These two simple components are at the basis of the proposed definition of surprise below.

The background information of an observer is captured by his/her/its prior probability distribution $\{P(M)\}_{M\in\mathscr{M}}$ over the hypotheses or models M in a model space $\mathscr{M}$. At a high level of abstraction and for, e.g., a human observer, the ensemble $\mathscr{M}$ may for instance consist of a number of cognitive hypotheses or models of the world, such as:

$$\mathscr{M} = \{\text{it will rain tomorrow; the cold war is over; the USC-Trojans football team is on a winning streak; my wallet is in my possession; my car is in good working order; my credit card information is secure; nobody at work knows that today is my birthday; etc}\} \quad (1)$$

At lower levels of abstraction and for less sophisticated observers, the model space may be much simpler, corresponding to straightforward hypotheses over well-defined quantities, such as, for example, the amount of light hitting a given photoreceptor:

$$\mathscr{M} = \{\text{light level is low; light level is medium; light level is high; etc}\} \quad (2)$$

With each of these hypotheses or models M is associated a likelihood function, $P(D|M)$, which quantifies how likely any data observation D is under the assumption that a particular model M is correct.

Given the prior distribution of beliefs before the next observation of data, the fundamental effect of a new data observation D on the observer is to change the prior distribution $\{P(M)\}_{M\in\mathscr{M}}$ into the posterior distribution $\{P(M|D)\}_{M\in\mathscr{M}}$ via Bayes' theorem, whereby

$$\forall M \in \mathscr{M}, \qquad P(M|D) = \frac{P(D|M)}{P(D)}P(M). \quad (3)$$

In this framework, the new data observation D carries no surprise if it leaves the observer's beliefs unaffected, that is, if the posterior distribution over the ensemble $\mathscr{M}$ is identical to the prior. Conversely, D is surprising if the posterior distribution after observing D significantly differs from the prior distribution. Therefore we formally measure surprise by quantifying the distance (or dissimilarity) between the posterior and prior distributions. Computing such distance between two probability distributions is best done using the relative entropy or Kullback-Leibler (KL) divergence (Kullback, 1959). Thus, surprise is defined by the average of the log-odd ratio:

$$S(D,\mathscr{M}) = KL(P(M|D),P(M)) = \int_{\mathscr{M}} P(M|D)\log\frac{P(M|D)}{P(M)}\,\mathrm{d}M \quad (4)$$

taken with respect to the posterior distribution over the model space $\mathscr{M}$. For example, using the premises of Eq. (1), if the data observation D consisted of patting your pocket and realizing that it feels unusually empty, that would create surprise as your posterior beliefs in the hypotheses "my wallet is in my possession" and "my credit card information is secure" would be dramatically lower than the prior beliefs in these hypotheses, resulting in a large KL distance between posterior and prior over all hypotheses, and in large surprise.

Note that KL is not symmetric but has well-known theoretical advantages, including invariance with respect to reparameterizations. A unit of surprise – a "*wow*" – may then be defined for a single model M as the amount of surprise corresponding to a two-fold variation between $P(M|D)$ and $P(M)$, i.e., as $\log P(M|D)/P(M)$ (with log taken in base 2). The total number of wows experienced when simultaneously considering all models is obtained through the integration in Eq. (4). In the following section, we provide a simple description of how surprise may be computed, and of how it fundamentally differs from Shannon's notion of information (notably, Shannon's entropy requires integration over the space $\mathscr{D}$ of all pos-